DOCUMENT RESUME

ED 354 333 CE 063 037

AUTHOR Peddie, Roger
TITLE Beyond the Norm? An Introduction to Standards-based Assessment. Developing a Qualifications Framework for New Zealand.
INSTITUTION New Zealand Qualifications Authority, Wellington.
REPORT NO ISBN-0-908927-21-5
PUB DATE 92
NOTE 53p.; For related documents, see CE 063 028-031 and CE 063 034-036.
AVAILABLE FROM New Zealand Qualifications Authority, P.O. Box 160, Wellington, New Zealand.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; Competency Based Education; *Criterion Referenced Tests; *Educational Testing; Employment Qualifications; Foreign Countries; Grades (Scholastic); Grading; High Schools; *Norm Referenced Tests; Outcomes of Education; Postsecondary Education; Standardized Tests; *Standards; *Student Evaluation; Student Improvement; Test Interpretation; Test Norms
IDENTIFIERS *National Qualifications Framework (New Zealand); *New Zealand

ABSTRACT

Standards-based assessment is the student evaluation method favored by the National Qualifications Framework developed by the New Zealand National Qualifications Authority. Before determining an assessment method, definitions of key terms and concepts such as assessment, validity, and reliability must be determined. Good assessments are developed more easily and effectively when assessors have clear purposes and a clear understanding of the strengths and weaknesses of different types. Norm-referenced assessment compares the results each learner achieves with what other learners achieve on the same test. In standards-based assessment, the outcome is analyzed against some fixed criterion or level of achievement. One subtype, competency-based or criterion-referenced assessment, sets a particular standard of competence that must be reached to receive credit. In another subtype, achievement-based assessment, a number of progressively more demanding standards are used and learner achievement is reported in the form of a grade. Critical issues in standards-based assessment center on the following: (1) theory versus practice; (2) how many and what type of assessments are needed; (3) awarding of merit; (4) test difficulty; and (5) test bias. Two extended examples of this kind of assessment are provided; the first concerns assessment of listening in a foreign language unit and the second concerns assessment in a unit on sales techniques; these examples illustrate the need for careful consideration of purpose, validity, reliability, and practicality. Educators should be clear about purposes, choose an appropriate form of assessment, then select the most valid and reliable measures that are usable in practice. (CML)

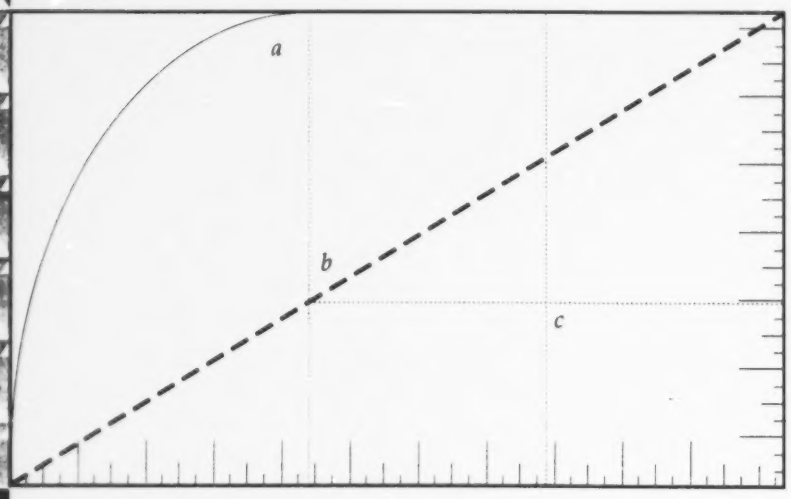
ED354333

DEVELOPING A
QUALIFICATIONS FRAMEWORK
FOR NEW ZEALAND

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A Baker

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."




# BEYOND THE NORM?

## An Introduction to Standards-based Assessment

*Roger Peddie*

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

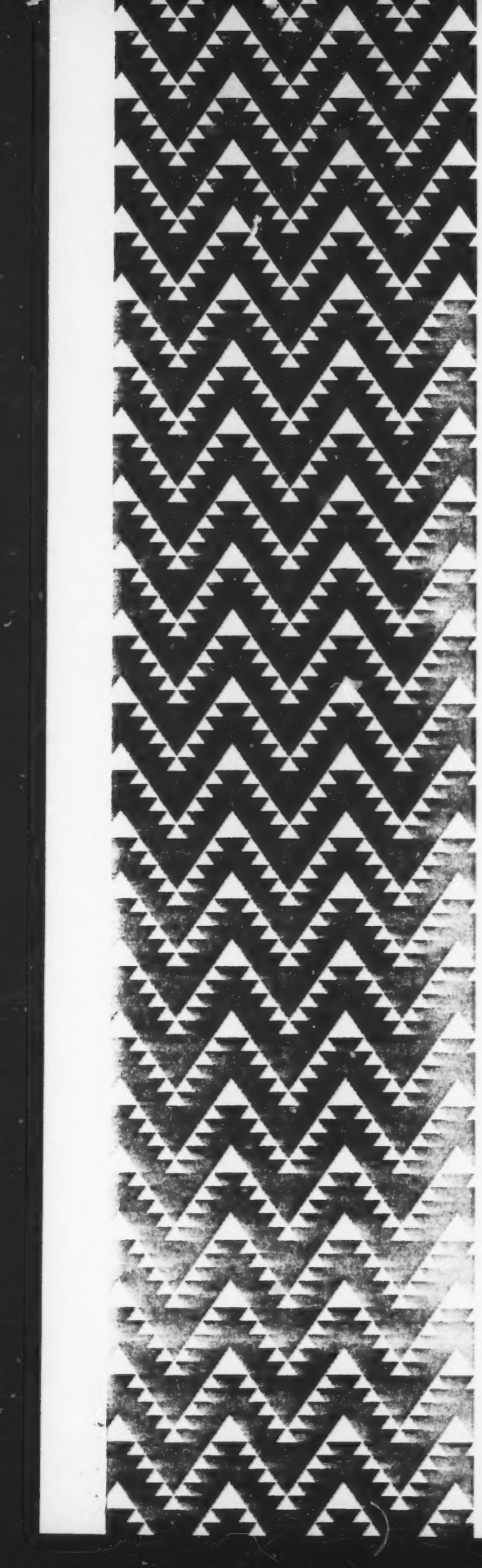
☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

NEW ZEALAND QUALIFICATIONS AUTHOURITY
Mana Tohu Matauranga o Aotearoa



BEST COPY AVAILABLE



**The New Zealand Qualifications Authority will promote improvement in the quality of education and training in New Zealand through the development and maintenance of a comprehensive, accessible and flexible National Qualifications Framework.**

*The Authority's main functions are to:*

- coordinate all qualifications in post-compulsory education and training (from upper secondary to degree level) so they have a purpose and relationship to one another that the public and students can understand
- set and regularly review standards as they relate to qualifications
- ensure New Zealand qualifications are recognised overseas and overseas qualifications are recognised in New Zealand
- administer national examinations, both secondary and tertiary

**© New Zealand Qualifications Authority 1992**

All rights reserved. No part of this publication may be reproduced by any means without the prior permission of the New Zealand Qualifications Authority.

Ruia taitea

Ka tu te kaikaha anake.

Strip away the sapwood

Let the heartwood remain.

# PREFACE

This was not an easy booklet to write, in part because so many people offered so much to me: papers, documents, advice, suggestions and even the occasional attempt at a directive. There is a vast literature on assessment, and much has been written in New Zealand on the subject over the last few years. The more I read, the more clearly the limits of a booklet like this became apparent. Nevertheless, I hope that, within the boundaries defined in the introduction, this booklet will be prove useful.

My thanks go to the many people who have advised me during the production of this booklet. In particular I am grateful to Don McAlpine for his assessment expertise, and to the Research Unit for Maori Education at the University of Auckland for their guidance. A number of people from teaching, private training, and government agencies gave freely of their time and knowledge. My thanks go to them all.

The views expressed here on assessment are those of the author, and are not necessarily shared by my colleagues, nor by the University of Auckland Administration. Indeed, given the eagerness with which some academics (myself included), seize upon every new document emanating from government education agencies, it is with more than the usual reserve that I accept responsibility for the final version of this booklet. If, however, it helps to generate some questions in the minds of teachers and some understanding in the practice of those developing new curricula, I will feel my time has been well used.

Roger Peddie.

**Roger Peddie**

# CONTENTS


| | | |
|---|---|---|
| | Preface | 2 |
| 1 | Introduction | 4 |
| 2 | Some Statements About Assessment | 6 |
| 3 | Definitions: Terms We Love To Hate | 10 |
| 4 | Standards-Based Assessment | 21 |
| 5 | Some Critical Issues | 29 |
| 6 | Assessment Examples | 36 |
| 7 | Concluding Remarks | 48 |
| | Further Reading | 49 |


November 1992

# INTRODUCTION

This booklet is an **introduction to assessment**, and in particular to standards-based assessment. More specifically, it is an introductory guide to assessment for units registered in the new Qualifications Framework. Although the focus is on standards-based assessment, readers should note that this is only **one** method of assessing student learning. The method is, however, an important part of the Qualifications Framework developed by the New Zealand Qualifications Authority.

The booklet also **introduces selected issues in assessment**, including some where there are unanswered questions. It is written for teachers, tutors, trainers and course developers[1] in institutions involved in education or training. An important gauge of its success will be if readers become more aware of problems in assessment, and learn never to be satisfied with their current assessment programmes.

The issues and examples offered in this booklet are from a limited range of subject areas. Teachers may find it useful to think of examples from their own subject/content area as they read each section.

This is **not** a textbook on assessment, nor is it a simple cookbook for producing assessments. It does not treat key ideas like validity and reliability in depth or detail. It does not cover statistical measures in testing at all. Nevertheless it does cover some of the required knowledge for good assessment.

The booklet does not deal with recording or reporting of test results. Several other topics are touched on only very briefly. These include the awarding of **merit**, and **moderation**, the monitoring of assessments by more than one teacher or institution to see if standards are comparable. Further discussion on these issues is found in other booklets.

This present booklet is not simply the "official" view of assessment. It does try to explain standards-based assessment, the form of assessment the Qualifications Authority is advocating in the Framework. It also notes some problems in assessment. The existence of these problems reinforces the point that experts in this field have long understood; **there is no single "correct" way to assess student achievement.**

Finally, while this is only an introduction, almost every page contains issues which require a good deal of both thought and action if they are to affect the ways in which teachers assess their students.

*1 Henceforth as a matter of simplicity, the term "teacher" will be used for all involved in education and training.*

## HOW THIS BOOKLET IS WRITTEN AND ORGANISED

The booklet is organised in a number of brief sections. Every section introduces some problems and issues. Readers should remember that even if the specific problems discussed later in the booklet can be solved, these solutions are often undermined by problems raised earlier. Yet careful attention to these problems should help to create better assessment programmes.

**Section 2** introduces a selection of twelve key statements about assessment. These raise issues which teachers should be aware of **before** developing an assessment programme.

**Section 3** defines and discusses some important terms used in assessment. It also introduces in a non-technical way the key issues of **validity** and **reliability**.

**Section 4** stresses the need to be clear about the purpose of assessment. It then introduces **norm-referenced assessment** before looking more closely at **standards-based assessment**. The focus is on two main types, **competency-based** assessment and **achievement-based** assessment.

**Section 5** gives a brief outline of five critical issues facing those developing assessment programmes: practicality, numbers of tests, merit, test difficulty, and bias.

**Section 6** consists of two extended examples, one from general (school) education and one from industry/training. In each example, however, issues are raised which are relevant to teachers in **both** fields.

**Section 7** offers a few concluding remarks.

The decision was taken early to make this booklet as readable as possible. This in turn led to the decision to write less formally and to omit references. There is a brief **Further Reading** section at the end, and it should be stressed that there is little or nothing in this booklet that cannot be found in standard texts.

The brevity of the reading list is deliberate. Learning about assessment by reading is an important but limited way to improve practice. Substantive and quality professional development programmes are needed to give opportunities for teachers to get practice in assessment, and to discuss and reflect on that practice.

# 2 SOME STATEMENTS ABOUT ASSESSMENT

This section presents a personal selection of twelve important statements about assessment. These do not cover everything that could be said, but what they do cover is important. The ideas offered here should be considered carefully by teachers developing assessment programmes. It also is worth coming back to a list like this after the programme is developed.

**1 Assessment should be as fair, accurate and appropriate as possible.**

Other words could be used here, like valid, reliable, equitable, consistent and "fit for purpose", but the message will be the same.

These things should **never** be "taken as read". Assessment affects human beings and their lives, a point which teachers should always keep in the forefront of their thinking. Developing assessments which are fair, accurate and appropriate requires time, expertise and resources.

**2 A good assessment programme is always an integral part of a good curriculum.**

In other words, assessments should never be developed separately from the rest of the teaching-learning process.

Assessment **policies** for national qualifications are now developed by the New Zealand Qualifications Authority. They are the body legally appointed to make these policies and to oversee the setting of standards for national qualifications. Teachers, however, still have important decisions to make in assessment. One of these is to ensure there is no split between the curriculum, teaching and assessment.

What this means in practice is that when teachers plan the "delivery" part of a unit in the Qualifications Framework, they should always integrate the planning of content, teaching and assessment.

The only case where assessment seems not to be part of the curriculum is when a learner is tested to determine recognition of prior learning. But a moment's thought will show that the assessment should still clearly relate to the curriculum goals of the course or programme.

**3 A good assessment programme should encourage and assist learners.**

This is more likely to happen when the assessment programme uses appropriate methods with realistic standards.

The clearer learners are about what is expected, and the more encouragement they receive to reach their goals, the less any assessment programme will be seen as threatening or negative. This of course involves not only the assessments, but the ways in which the teacher teaches. Nevertheless, clear assessment procedures, equally clear instructions about what is required, and a known and reachable standard will always be helpful to learners.

**4 Most assessments involve value judgements and are intertwined with ethical issues.**

Even if we assess against apparently quite objective or neutral "standards", an assessment involves an evaluation or judgement about a performance. The results of our assessment programmes often categorise people in ways that society values or rejects.

If Mary "passes" a driving test and John "fails", then Mary can legally drive a car. To suggest that this is completely neutral is incorrect. John's attitude and beliefs about his ability, and the attitudes of those who know him are clearly involved here, and the same is true of Mary. Whether this is a good thing or not does not really matter; it happens, and it happens as a result of assessments in schools, post-school institutions and in everyday life.

**5 Assessment programmes should be as unbiased as possible. In particular, they should not be biased against identifiable groups in the community.**

Assessment is always "framed" by the dominant society and culture. Potentially therefore, some groups can be disadvantaged even before any assessment takes place.

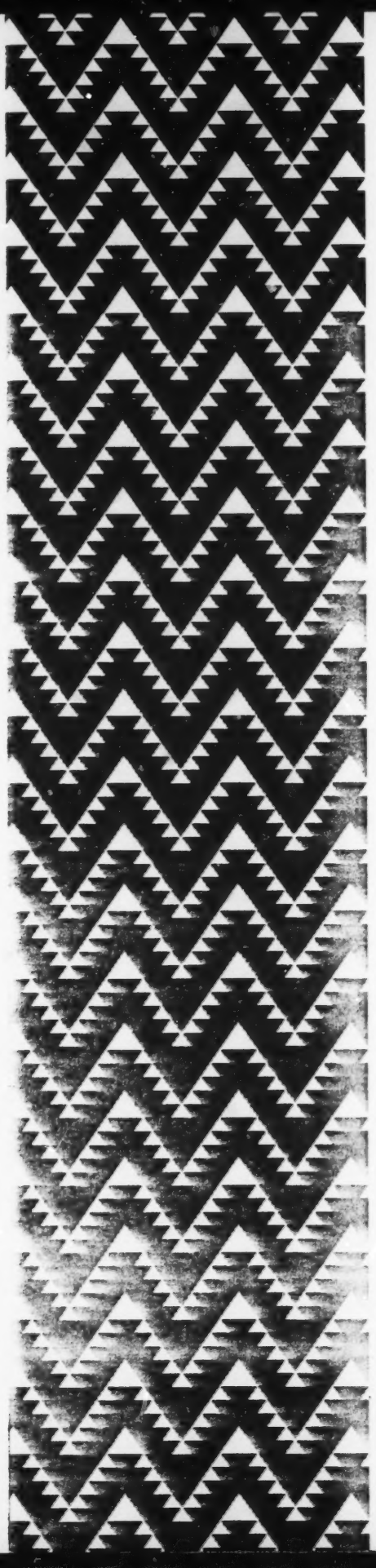
Perhaps the most obvious examples relate to Maori, to women, and to cultural and ethnic minorities. For some tasks, assessment which is competitive, individualistic, written and lacks any spiritual element will disadvantage some Maori. Assessment which uses sexist language, has male-dominated examples and is based on male notions about how things should be done will similarly disadvantage many women (see also Section Five).

**6 Teachers should be as aware as possible of what comparisons and standards are involved in their assessments.**

It can be argued that all assessments involve comparisons, just as all assessments involve setting up some form of "standard". Awareness of this can help to avoid unrealistic test standards or standards based on faulty expectations.

Suppose we are assessing some learners to see if they can perform a very simple skill, where there are no comparisons or ranking intended. Even in such a case, we probably had an underlying notion about what the average - or good - student could be expected to do at the time of the assessment. In other words, we assess in part on the basis of some conscious or subconscious expectation about the way students should perform.

With tasks that are quite complex, this is perhaps more obvious. A fair and conscientious traffic officer may judge a candidate on a driving test against quite clearly defined standards. But these standards reflect what we think a newly licensed driver ought to be able to do in order to drive safely. They were set up within a range of "driving skills", classifying drivers from menace to expert. Learner-driver "competence", therefore, is assessed against a background of comparisons with drivers who have licences, and with those who should not be allowed on the road.



**7 Assessment implies some form of subjective analysis or judgement, both at the time of preparing the tests or measures to be used, and after results have been obtained.**

There is no such thing as an "objective" test. We choose what to test, we choose when and how to administer a test, we choose what messages to give and to whom these messages should go after the test.

Of course, some assessments **are** more objective than others. For example, we can **score** a test objectively. A computer-marked multi-choice test will give the same results each time we feed the answer sheets through. This is certainly **not** the case with essays, and there are serious questions about the objectivity of assessments which involve observation of performance (see later).

**8 Decisions about "passing" and "failing" almost always involve arbitrary and subjective judgements.**

This is an important extension of statement seven.

First, there is no **logical** reason why a mark of fifty per cent should mean that a learner "passes" an examination while a mark of forty-nine per cent should mean another learner "fails". As those familiar with statistics will know, such a difference regularly falls within the standard error of measurement.

Even when we set up "assessments of competence" (see later), the level of achievement we take as acceptable is unlikely to be driven solely by logic or by a completely objective decision.

Most of us probably accept that assessing a learner's ability to write a literary essay in English or Maori is somewhat subjective. But assessing whether a learner has the ability to use a lathe involves subjective judgments too. In this case it is simply more likely that we decided in advance on the precise conditions necessary for the learner to demonstrate what we count as "competence".

**9 Assessment can be used for a number of different purposes. Some approaches to assessment suit some purposes better than others.**

Another way of saying this is, Be sure to decide on your purpose for assessing **first**. Then choose the **most appropriate** method for this purpose.

If we are using assessment to select the "best" candidate for a job, then we often want to compare the candidates with each other. This is especially true if a number of them come equipped with the formal qualifications we require. Some way of ranking the candidates then becomes important.

If we are trying to improve our teaching programme, then we need to evaluate the **programme**, and not just assess the students (cf. Section Three). This is often best done during the course of our teaching, and will not involve ranking students at all.

If we want to see whether students can do what we have taught them to do, we often set up an assessment programme at the end of our teaching, and assess them to see if they **can** perform at the appropriate level.

It is quite possible that the "tests" we use may look quite similar in these different cases, but the approach, the advice given to the students and, above all, the interpretation of the results will differ.

**10 There are ways of deciding whether particular tests or ways of measuring are "better" or "worse" than others.**

What this means in practice is that "good" tests produce more valid, reliable and usable results than "bad" tests. We will discuss validity and reliability in Section Three, but an example here will introduce and perhaps clarify some of the issues.

If we set an essay test when our aim is to see if a student can cut hair, we find out if a student **knows** how to cut hair. To test what we in fact want to test, we should ask the student to cut someone's hair. If, however, we then choose as an assessor someone who does not know exactly what the course requires, we have a poor assessment for a different reason. The key point here is that we should learn as much as we can about assessment so we can improve our own assessment programmes.

**11 The "technical bits" of assessment do make a difference.**

This follows on from the previous point and may seem too obvious to say. Yet we all need to be reminded regularly of this point.

The way the teacher sets a test, organises test procedures and the ways in which scores are allotted, combined and reported all have significant effects on the eventual outcome for individual students.

To take a simple example: if we decide to allow students six hours instead of three in an accountancy examination, a number of students may reach the desired standard who would otherwise not have done so. If speed was a part of what we were looking for, changing the time allowed will clearly change the standard and affect a number of students' results.

Similarly, suppose we combine the scores from twelve different classroom tests without being aware of the effects of such things as test difficulty and the range of scores in each test. If we then award "merit" to students who score above a certain level, our selection may be a reflection of the way we did the calculation. We may not actually have selected the most deserving students.

**12 Even experienced assessment specialists need to go back to the basics every time they are involved in developing a new assessment programme.**

Because no assessment system is perfect, we need to keep challenging our own practice, and to remain aware of the issues raised by these statements whenever we assess.

# 3 DEFINITIONS: TERMS WE LOVE TO HATE

Some terms used in assessment seem to be particularly disliked by many teachers. This is in part because they consider them to be of minor importance or even irrelevant to classroom practice. In preparing for this booklet I was told by one adviser:

> Of course, you **have** to talk about validity, but don't mention it by name or you'll turn teachers off. They can't be bothered with the jargon of testing and assessment.

However, informed discussion of many of the issues really requires the understanding of a number of key terms. This understanding is essential if teachers are to read more widely on some of these issues.

This does not seem to be a problem in the subject areas in which teachers work. Geographers throw around terms like "isobar" or "hectopascal" without a moment's thought; language teachers discuss the relative merits of "audiolingual" and "communicative" programmes; while tutors in automotive engineering know all about "carburettors" and "fuel injection".

Yet there is an issue here. The language of assessment can cause problems, in part because different writers on assessment sometimes use similar terms to refer to slightly different things. The other thing they do is to define their terms in rather technical and precise language.

What this section aims to do is to introduce as clearly as possible some key terms needed to discuss assessment. It is accepted that this means that the precision some might expect will not be found here; this has been a deliberate decision in line with the intentions of the booklet. Note that discussion of various **types of assessment** is found in the next section.

## ASSESSMENT

Some writers make careful and precise distinctions between "testing", "measuring", "assessing" and "evaluating". Others treat these as almost identical. Generally speaking, however, there is quite a lot of agreement over what follows.

First, a **test** can be any way in which you find out more about a student's ability, knowledge, skill or understanding. It usually implies a controlled situation, like a School Certificate examination, but there is no reason at all to limit the notion of a "test" in this way. Indeed, there are good reasons for saying many tests should consist of tasks carried out in the normal way in which the person is going to perform.

Some tasks are already like this. The final and most important test of driving to gain a full licence is not carried out on a closed circuit, but on the kind of roads where it is assumed the learner will later drive.

A test usually comprises one or more **items**. We usually associate the word item with a single question in a short-answer or multiple choice test, but it is useful to think of every separate part of a test as an item. So, if a test of ability in sewing involves the making of three different garments, the making of each garment is an "item" in this test.

**Measurement** is usually taken to refer to the way in which we count, grade, or give a score to someone, **with nothing more implied**. In other words, it is usually used to refer just to the way we arrive at a result, and not any judgement about how good or bad that result is. We need a context or frame of reference to move from a measurement to an assessment.

An **assessment** is normally taken to occur when we make some analysis, or offer a description of the measurement we have taken. If a learner tells you they scored a D on a test, what happens? We automatically assume they have done badly, not because the letter "D" has some magic property, but because we know a D is awarded for marks below the pass/fail line.

The **measurement** (D) becomes an **assessment** through our understanding of its meaning. We are analysing the measurement that was made and interpreting it as "bad".

Notice that, to assess in this way, we need to know what scale is being used (often, A to E, or 100 to 0 percent), and we are often interested in how well other students did on the test. But that is partly because in this case all we were given was a letter of the alphabet.

If the same learner reported they had passed their driving licence, we would not normally need to ask for more information. We might ask what the instructor was like, and how the student felt about taking the test. The result itself would not need any further interpretation. This is because we know what is required for a driving test. We also know that a learner does not pass depending on what happens to other novice drivers on that day.

A second and broader meaning of assessment is the whole process of collecting evidence about what a learner has achieved and the interpretation which follows. This is what teachers mean when they talk about "assessing" their class's ability in mathematics.

**Evaluation** tends to be used to imply some form of judgement, either of students or about programmes. There **are** some forms of evaluation which involve little or no judgement, but which involve the whole learning community in a positive form of programme development. In this booklet, evaluation of programmes is not discussed, and the term evaluation will not be used after this.

Two terms which commonly turn up in discussion on assessment are **formative** and **summative**. These two terms are dealt with here rather than in the next section because they are not actually **types** of test. Rather **they refer to two different purposes for which tests can be used.**

**Formative assessments** refers to measures taken during a course or programme which are aimed primarily at giving feedback to teachers and students. Such measures are commonly used to find out how much has been learned so far, or how the course is actually going, or what might be done to improve the learning process.

When the aim of the formative measure is to diagnose current learner needs, or to evaluate how successful the teaching or course is to date, it would be inappropriate to use the results as part of the final grade. Sometimes, however, a test used to gather information does this by measuring performance on one or more of the learning outcomes. In such cases, teachers may well use the results both to get/give feedback and as part of the final grade. When this happens, we often refer to it as **oncourse assessment** even though one purpose of the test is formative.

**Summative assessments** are those based on tests given at the end of a course, or section of a course. The information they provide is usually the basis for the reported grade, level of achievement, or statement about whether the student has reached the minimum standard for a "pass". In the Qualifications Framework, summative assessments may be used at the end of a unit to see if learners have reached the agreed criterion level for credit in that unit.

Finally here, **moderation** refers to the act of checking an assessment in some way to see whether that assessment is roughly the same as comparable assessments by other teachers or at other institutions. When, for example, another teacher at the same institution re-marks a test we have a form of **internal moderation**. When a sample of work is sent to an outside person or body for checking, this is a form of **external moderation**. A companion booklet on moderation, prepared by the Qualifications Authority, deals with this in much more detail.

## VALIDITY

Validity is one of the terms teachers love to hate above all others. A definition is in fact fairly easy to understand. A **valid test is one which actually tests what it sets out to test, and not something else.**[2] Another simple way of thinking about validity is that the assessment is "fit for purpose"; it is an appropriate way to assess the learning outcomes you want to assess.

If I measure your height with a ruler which is marked as a metre long, but which is actually only ninety centimetres, then I will never get a valid measure of your height, no matter how carefully or how often I measure.

If a Kaumatua wants to judge your ability to speak on a marae, it would be not be a valid test of that ability to ask you to write down what you might say. But note that this **would** be a valid test of whether you **knew** what to say in such a speech.

2 *Strictly speaking, validity and reliability refer to the test* ***results****, but developing valid and reliable* ***tests*** *will produce results we can use with confidence.*

Most teachers understand that. Some also understand that there are a number of different forms of validity: for example, **face validity, content validity, construct validity and predictive validity**. While it is useful to know about all of these, this booklet will deal briefly with face validity and then focus on content validity. Content validity is the most important for many teacher-developed assessment programmes.

Before discussing these types of validity, however, it is worth really stressing one critical point:

***an assessment which is not valid is a waste of time.***

It is also unfair, unproductive, and potentially destructive. Students who fail because a test is invalid may simply give up. This may easily affect their career choices and their future lives. It does not matter how **reliable** an invalid assessment may be, it is still a waste of time so all teachers **need** to understand validity, and they need to understand it well.

## Face Validity

Face validity relates quite simply to how the test looks. Does it look like a test which measures the achievements and outcomes which interest us? If the answer is 'yes', then the test has good face validity. While in one sense this may be thought unimportant, face validity plays an important role psychologically. It helps students and the community to accept that a test is doing what it should.

An example may help here. A school may run a competition in listening and speaking in a foreign language. A native speaker is brought in to run the competition and judge the winner. The students might normally expect that such a competition would involve conversation with the native speaker. The native speaker could decide instead to give a series of sub-tests of such things as pronunciation, knowledge of vocabulary, listening abilities and the ability to respond to questions.

Such an assessment would almost certainly lack face validity for the students and probably for some of their teachers. This in turn may put some students off, or result in performances that were not typical simply because students did not think the way the test was run was how they expected it to be.

So, to sum up here: a test ought to **look** as though it is valid, as well as **be** valid.

## Content Validity

In teacher-made tests, **and** national examinations, content validity is the most important aspect of the assessment. If an assessment or test has high content validity then it fairly and faithfully reflects the learning the unit is aimed at producing. Despite the name, it is not just a question of "content" in the narrow sense of "the stuff they learn". If the goal of the unit or course is to produce skills, then a test of skills is needed to ensure content validity.

Note that **we judge the validity of a test in a particular context and for a particular purpose.** This is important. We can never say that a test is simply "valid", full stop. The same test can be valid for one assessment but not for another.

If the learning outcome of a unit in home economics is **knowing** what items need to be included in an annual home budget, then a valid test is to ask the student to tell you what these are. The same test would be a much less valid way of assessing whether students could actually **produce** for their own home a budget which would keep them out of debt.

Next, **each test** contributes to the overall validity of results in a unit or course. This is one of the main reasons why a course or unit should be planned as a whole. Whether we call them "aims", "elements", learning outcomes, or any other name, **the goals of a course should be reflected in the course content, the way the course is taught, and in the measures used to assess learners on the course.**

There is nothing too complex about the **idea** of content validity. The problem is how to ensure that we actually plan, construct and carry out valid assessments. This underlies much of the discussion in the remaining sections of this booklet.

The examples which follow are aimed at demonstrating the importance of two related points.

- It is usually not a question of test results being "valid" or "not valid", but of trying to **increase** the validity within the practical constraints of a programme.
- There are always **more** and **less valid** ways of assessing any type of learning outcome.

### Example 1

One element or learning outcome in a horticulture unit relates to the ability to choose appropriate grass seed for a new bowling green. An assessment with some validity would be to ask students to identify the grass on an existing green, and to say whether it was an appropriate type. A more valid assessment would be to ask students to write down which grass they would buy for greens in different local areas. An even more valid assessment would be to take them to the site of a new green, then to a garden store, ask them to buy appropriate grass seed and to explain their choice.

At this stage, we have not considered practical issues in assessment. This example and the next two raise some of those issues (cf. "Theory and Practice" in Section Five).

### Example 2

A learning outcome in an early unit in Te Reo Maori relates to understanding

basic greetings and responding to them appropriately. An assessment with low validity would be to give a series of written greetings and responses in cartoon form and ask the students if the respondent replied appropriately. An assessment with higher validity would be to give some greetings in writing and ask the students to write what they would say. Better still would be to **say** a series of greetings and ask the students to write their response.

It may seem that the assessment with most validity would be to test the students individually, saying one or two greetings and seeing if they responded appropriately. In fact, an even more valid way of testing would be to take the students to a hui or onto a marae and to observe the ways in which they responded to greetings. Again, issues of practicality are obvious.

### Example 3

An employee in a bank takes a unit on the proper way to supply and maintain ATMs (those machines that give you money when the banks are shut). An assessment with very low validity would be to ask the employee to write down what they would do.

An assessment with higher validity would be to ask the employee to demonstrate to the trainer all the required tasks which formed the learning outcomes for the unit. An assessment with even higher validity would be to monitor the employee's work performance to see if they actually **did** what they had learned in the unit.

A further consideration is worth noting. Valid results depend not only on the test itself, but also on what we decide to mark or "count for success". This is different from marking accurately and consistently (see the next section).

If we are assessing the ability to interpret a poem, it would be invalid to give low marks to people who made lots of spelling mistakes which did not affect the meaning - even if this annoyed us intensely. In the same way, it would be invalid to penalise someone for turning up to an engineering test in what we felt were untidy clothes. But it **would** be valid to take clothing into account if the test involved machinery, and the learner's clothers could be caught in a machine.

One final point must be stressed: validity is so important that it is worth spending as much time and effort as we can to make our assessments as valid as possible.

## RELIABILITY

Broadly speaking, **reliability refers to the consistency with which a test or assessment programme measures what it sets out to measure.**[3] It does not imply that the test is appropriate (valid), but that the results can be "relied on" to give an consistent and accurate picture of what we measured.

*3 As noted under validity, reliability in fact relates to the **scores** produced by a test or measurement, not to the test itself. Saying a test is "reliable" actually means that it produces reliable (consistent) **results**.*

There are several forms of reliability, some of which can involve quite complex statistical measures. As this is an introductory booklet, we will not be considering statistical measures here.

By the way, there **are** links between reliability and validity. A test needs to be reliable to have a high degree of validity, even though reliability by itself does **not** guarantee validity.

Sometimes, for example, a reliable instrument still gives us the wrong information. Suppose I have a stopwatch which actually runs three seconds fast every minute, and I am timing a chemical reaction which a learner has to set up to take place in exactly one minute. If the learner sets the experiment up so that it takes sixty-three seconds, I will believe the learner has reached the standard set. My watch will show only sixty seconds. I can ask the learner to do it ten times, and each time the watch would show me the same result.

Believe it or not, in that situation we would have a reliable measure. The watch measures consistently and "accurately" each time. It just so happens that the apparent accuracy is incorrect. So, what we have here is a reliable but **invalid** instrument. The watch is not measuring what we set out to measure.

Reliability does not just depend on accurate equipment. There are all sorts of chance factors operating in a test situation. Because of this, we should often think seriously about assessing more than once. Trying to avoid chance factors is especially important when the outcome of the test is really critical for the student. An example would be if the student were taking a final examination to qualify for an award or career. A lot of attacks on School Certificate were over the fact that each examination tested the student on a single occasion.

On the other hand, there are some tests where a single failure to reach competence may be enough to make us think again. If the learning outcome involves the application of anaesthetic in exactly the correct manner to avoid serious problems, we clearly want the learner to demonstrate competence **every** time!

To improve "test" reliability, we try to ensure that:

- the test is long enough to reduce inconsistency due to chance factors;
- the test consistently measures in the same way over time with equivalent groups of learners;
- where it is appropriate, the test will assess a random sample of behaviour, or be made up of items which randomly sample what the learner had to learn;

- where appropriate, the test is "internally consistent", testing equivalent things in each of its parts;
- where we have two versions of a test, the results produced are about the same for each test.

Each of these aspects will now be considered briefly.

## Test Length

A very short test is more subject to chance results than a longer one. This is fairly easy to see. Suppose in a course on engine maintenance there are a large number of specific points which the learner must know, but I decide to test only ten items. The results of my test are likely to include some quite inaccurate scores.

Some students may have a very good general knowledge, but my test quite accidentally selected several of the things they did **not** know. I may also have picked items which were the **only** ones which one or two weak students happened to know. They will score better than their "true" ability. If I want a test which will consistently (reliably) select those who really **do** know a lot, a test with more items will be a good starting point. In this case, a longer test would also tend to give a more valid picture of their knowledge.

There are statistical ways which help to show what random variations we might expect in some kinds of test. Those developing assessment programmes might like to follow this up by further reading.

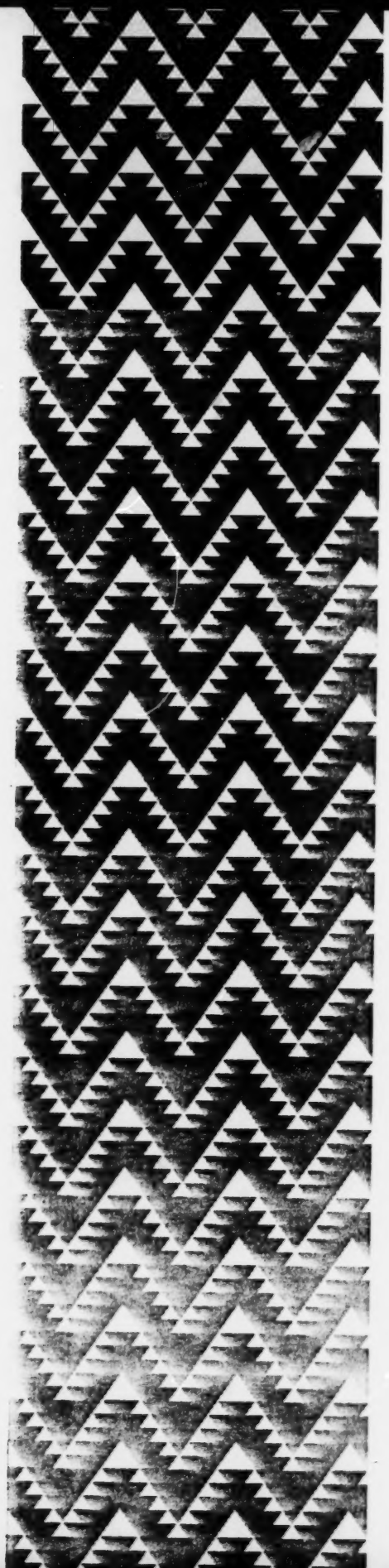
## Consistency Over Time

If we test behaviour of one or more groups of learners on two or three occasions and get the same results, this suggests our test has reasonable consistency over time. If we get a variety of results and we are reasonably sure the learners have not changed or learned more, we need to examine our test to see if it is unreliable.

Possible reasons include:

- the test is very short, and is subject to a lot of chance variation (see above);
- the instructions for administering the test are too vague, and so we are actually testing differently each time;
- some or all the test **items** are inaccurate, unclear, far too hard (or easy), or biased.

The need for tests to be reliable over time is increased when learners who initially are unsuccessful have the opportunity to take a test on a second or third occasion. This is one of the intentions of the new Qualifications Framework.



## Random Sampling

If our test is made up of a number of items, selecting a random sample from the total pool of items is quite a good way to improve reliability. It will help to eliminate some forms of bias, which can occur if we deliberately choose our set of questions or items. This is because our choice can easily match the expectations of some students but not others.

Of course, in some assessments, the students will know in advance exactly what items are included, either because these are absolutely **essential** for competence, or because it is **a skill or behaviour** that is being assessed. In this latter case, we would try to ensure that the assessment is made under fair conditions. The discussion of reliable observation later in this section offers some thoughts on this.

## Internal Consistency

We sometimes try to make tests more reliable by making sure items are consistent with each other. This is normally true when the items are all meant to be testing the same thing, or occasionally when all items are to receive the same mark or credit. In such cases, we also want each item to be equivalent in the results it gives.

Again, there are statistical formulae for checking on internal consistency.

## Test Equivalence

Sometimes we need to develop two different versions of the same test. This is likely to be quite common in some areas, as the Qualifications Framework encourages teachers to promote success by allowing students to take tests on more than one occasion. In this case, we need to be sure that the results students get are roughly the same, whichever version of the test they take.

This can be done informally by inspecting test results, or more formally by statistical measures. You can check on your own "parallel" tests informally by giving them to the same group of students and comparing results. If the scores are very closely equivalent, you can feel reasonably confident that the two tests can be used to give reliable (in this sense!) results.

## Other Factors

Like validity, we can make tests **more** reliable by being aware of the threats to reliability and trying to eliminate them. We aim, therefore, for an assessment which really does test consistently and fairly each time we give it, and that is not affected by things like "question-spotting" - picking a few likely questions and studying answers only to these questions.

Sometimes **testing in a variety of conditions** is part of an assessment. The learning outcomes for a unit on green-keeping may be tested quite deliberately in different weather conditions. The test results, however, would be **un**reliable if some learners were tested in the wet and others on sunny days, or in very humid conditions. By the way, testing in only one kind of weather would also be far less **valid**, assuming the aim of the unit was to produce good greenkeepers.

Secondly, **we need to make sure our scoring or measuring is as accurate and reliable as possible.** This is not the place to go into details, but it means we should be able to mark the same test on different days, for example, and come up with the same result. It means that two markers marking independently should also come up with the same result.

For this reason, a more objectively scored test, like a multiple-choice test, is usually more reliable than an essay, because essay marking is more likely to involve subjective judgements. Note, however, that there will be learning outcomes where an essay is definitely preferable as a **more valid** measure. In that case, we would choose to sacrifice some reliability, as validity is **always** more important.

For example, if the outcome of a unit in history is to show that the learner can develop a written argument in response to statements about defined historical events or periods, then an essay test is going to be more valid than even a complex form of multiple-choice or short-item test.

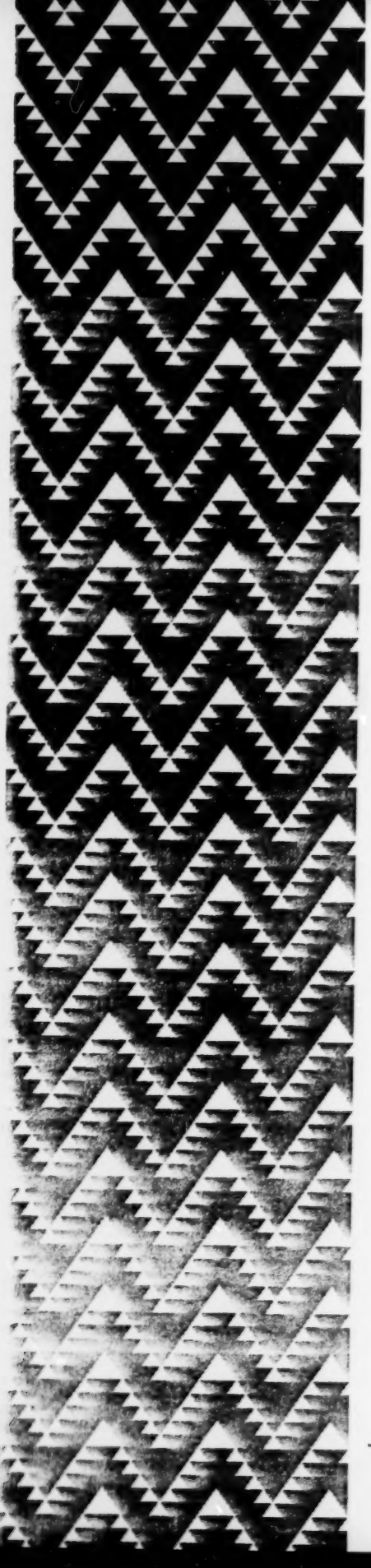
## *Reliable Observation*

Before leaving reliability, one further issue is worth exploring. It is possible that standards-based assessment will result in an increase of assessments using **observation of controlled or real-life performances**. While this may often be a valid and positive way to assess, teachers should be aware that reliable observation is a major area of concern.

It should not be thought that this relates only to casual observation, like witnesses giving different versions of a traffic accident. Even when you are (supposedly) concentrating, things can go wrong too!

Some of the difficulties associated with observation include the following.

- Deciding on - and obtaining - a fair and representative sample of the learner's behaviour.
- Seeing everything the learner is doing. This is sometimes difficult to do without being intrusive (e.g. peering over someone's shoulder during a complex repair). Sometimes seeing everything is simply impossible.
- Observing a number of things at once. This leads to unreliable recording and/or decisions.



- Accurate recording over time. Our attention drifts when we have to observe for any length of time. This becomes even more serious when we have to observe a series of learners doing the same thing.

These problems can be lessened by observing twice, or by having two independent observers recording and assessing a performance. Both of these options raise other questions, however, as most observed performances are going to be nerve-racking for learners (remember doing a morning talk at age six?).

It is better to try to improve the quality of observation and recording as we plan the assessment. We can try to arrange for some or all of the following.

- Plan for observation sessions which are the least threatening for the learner, so that the act of being observed does not affect their performance.
- Have a simple checklist for clearly defined behaviours to observe. Make sure this does not reduce the test to something that is no longer a valid measure.
- Record and/or make notes as soon as possible after observing, whenever recording cannot be done while you are actually observing.
- Use a second observer when this is non-threatening; otherwise consider the use of video for important or potentially dangerous tasks in which a high level of competence in realistic conditions is essential.
- Plan test conditions which take into account observer comfort, recording position, fatigue, etc.

## CONCLUDING REMARKS

This section has defined some key terms in assessment, and introduced the very important concepts of validity and reliability. Technical terms are one way of talking about a specialised field of knowledge. What is really important in assessment, however, is that we can **act** on our understanding and not just "know the words".

# 4 STANDARDS-BASED ASSESSMENT

Material published by the Qualifications Authority to date draws a clear distinction between two main types of assessment, **norm-referenced** and **standards-based**. Standards-based assessment is then divided into **competency-based** and **achievement-based** assessment. This section introduces these types of assessment, notes some similarities and differences between them, but concentrates on standards-based assessment.

Readers should note that there **are** other forms of assessment, as well as other ways in which people talk about assessment. In reading, you may come across such things as "domain-based" assessment, "objectives-based" assessment, and similar terms. These will not be discussed in the present booklet.

## ASSESSMENTS AND PURPOSE

Assessments are developed when we want to check how well people can perform.[4] If we believe that **no one** could achieve the goal we set, or if we believe that absolutely **everyone** could already achieve it, we would not bother to assess. The important point is this -

**good assessments are developed more easily and effectively when the assessor has a clear purpose in mind, and a clear understanding of the strengths and weaknesses of different types of assessment.**

When we decide to assess, we often have one of two purposes in mind:

- we want to see who can achieve set goals or standards after some learning or practice;
- we intend to use our assessments to rank or select among those who we are assessing.

Sometimes we want to do both.

Whichever way we go, we have in our minds some underlying standards, expectation and/or comparisons when we set up our assessment procedures. The clearer we can be about these things, the better chance we have of being clear about the form the assessment should take. We are also more likely to make clearer statements about the results we get.

Two examples may help illustrate this point. First, suppose we have an assessment in a language where the learner is judged competent if they can demonstrate reading with understanding. The teacher of a sixth form would presumably give a different passage to test this competence from the one which would be set for a third former. This is because they expect the older learner to read more competently, so the standard for reading with understanding is higher.

Secondly, suppose we have a procedure in a workshop where one task of a particular job is to turn off a safety valve if the pressure exceeds a certain

4 *"Perform" does not refer here only to skills and other observable behaviour; it includes demonstrations of understanding, knowledge and other forms of ability and learning.*

level. We would clearly want to assess this task as part of the procedures for that workshop.

This is because we do not have the absolute certainty that every worker **would** shut down the valve. Our assessment is to ensure that the person in charge will act as safety requires. We know it is **possible** to shut down the valve, but we test to make sure that successful applicants **will** shut it down.

## NORM-REFERENCED ASSESSMENT

In norm-referenced assessment, we are really comparing the results each learner achieves with what other learners achieve on the same test. We are sometimes comparing what each learner achieved with what was achieved by a large sample of learners, or "reference" group, on a previous occasion. That is, however, something of a simplification.

Perhaps the best known norm-referenced test in New Zealand is the School Certificate examination. A quick review of how the assessment process operated (until 1992), provides a useful illustration of some aspects of norm-referencing.

To start with, the examiner(s) set an examination which was checked by a moderator. Both people were trying to provide an exam which would do the following things:

- cover the main areas of learning as set down in the examination prescription for that subject;
- assess at about the same level as the year before;
- scatter the learners widely, so that the best could score 100% and the worst almost nothing.

After the examination was sat and marked, there were checks on markers, and statistics were sometimes used to adjust the marks of those found to be marking too hard or too easily. Then the marks were adjusted again using a sophisticated system to ensure that learners sitting that particular subject were getting the same sorts of marks as they were in other subjects.

In recent years, the scores were then reported in grades. In the "big" subjects like English and mathematics, the final results could be plotted onto a graph as a kind of bell-shaped curve.

One of the underlying assumptions was that this "normal distribution" as it is called, was a reasonable end product. In other words, one assumption was that abilities in each subject really are spread in this way.

Another assumption, strongly held by many people, was that such an examination was fair. This was partly because it did not depend on individual teachers, who might be biased for or against individual learners. Others accepted that the examination gave a reasonably accurate picture of what each learner could achieve.

A moment's reflection will show that the eventual score each learner received in a subject depended on a number of factors. One of these factors was the performance of **other** learners sitting that subject and other subjects in the same year. This and a number of related concerns have led to many teachers and some members of the public wanting a different system. They wanted each learner to be tested simply to see what **that learner** knew or could do.

There are a couple of very important points that need to be made before moving on. First, the School Certificate examinations had at their heart a **standard** which was, broadly speaking, related to an examination in which the average learner would score around fifty percent. This in turn was (and is) based on the examination prescription, which is developed from the curriculum. This curriculum is itself based in part on what teachers believe learners are capable of learning in the years leading up to the examinations.

Thus, though the focus in recent years has been on the inappropriateness of measuring learners against others, and on the "injustice" of scaling the marks actually gained, the School Certificate examinations have always related to standards.

Secondly, a well-conducted norm-referenced assessment programme **is** a sound way of assessing, especially if the goal is to select candidates according to a ranked order of merit. As a main purpose of this booklet is to look at standards-based assessment, this point will not be pursued. It is important to remember, however, that **the selection of an assessment procedure should always be made in terms of why you want to assess in the first place.**

## STANDARDS-BASED ASSESSMENT

The term **standards-based assessment** is used when the measurement or outcome is assessed, in other words, "analysed", against some fixed criterion or level of achievement known as a "standard". A whole set of standards may be involved. These standards should be set in advance, so that they are well-known to both teachers and learners. In theory, each learner gets exactly what they achieve, so that it is possible - again in theory - for **all** learners to achieve the particular standard desired.

The number of learners who actually achieve the standard(s) will depend on the level at which the standard is set. At times, of course, the minimum standard has to be set at a level which is needed for safety. In a unit which completes a qualification like building construction, for example, we would want to be sure that the minimum standard for credit would ensure that a house would not collapse on its occupants.

In many other cases, the standard is based on some expectation about what is achievable, and thus some form of comparison. One major difference is

that the expectations and comparisons should be based both on the prior experience of teachers and other experts, and a careful analysis of the unit and its learning outcomes. Neither the standards nor the final reported results depend on what a particular group taking the test happens to achieve.

Suppose we use a standards-based assessment for an advanced cooking class, and set the standard for one item as making a successful soufflé in less than 40 minutes. We will not worry about defining successful here, but note that such a definition would be critical! The **standard** set here includes a time limit, as in real life we do not normally have the luxury of spending days trying to get one soufflé just right.

The time allowed would ideally be decided upon by a group of experts who set a realistic standard based on their knowledge of what is possible and/or expected in a real-life situation. At the same time, this would be a standard that a good apprentice chef had some reasonable hope of reaching.

The clear intention of the Qualifications Framework is that realistic standards are set in terms of curriculum and of the desired learning outcomes specified in the Unit Standards. Thus, if the ability to strip a carburettor or develop a film in a certain amount of time is the specified learning outcome, then the standards will be set according to that goal.

In the same way, a business interpreting course in Arabic may have as one learning outcome the ability to interpret normal speech on commercial subjects simultaneously and without errors which affect the meaning of the original message. Here, the **realistic standard** has to be a high one, as the aim of the unit requires this. Nevertheless, all learners who reach this goal would be awarded credit for the unit.

Another shift worth noting is the Qualification Authority's intention that teachers and tutors provide opportunities for students to attempt to reach the standard on more than one occasion. This comes up in later discussion.

Standards-based assessment is used to cover a variety of types of assessment, but the two main sub-types referred to in the Qualifications Framework are **competency-based assessment** and **achievement-based assessment.**

## COMPETENCY-BASED ASSESSMENT

**Competency-based assessment is where we set a particular standard which candidates must reach if they are to be judged as "competent", and therefore receive credit for the unit of learning.** "Credit" here simply means the required level has been reached, and not the narrower meaning of "doing well". Students who **are** outstanding will qualify for **merit**.

The standard here, then is a **criterion** level in specified skills or areas of knowledge. This is why competency-based assessment is also sometimes known as **criterion-referenced assessment.**

If the learner does reach the required standard, our assessment is that they are competent to perform the task, or do or know what the unit is aimed at teaching. It does not imply the learner is "perfect", although in some units complete mastery of all elements may be necessary for health or safety reasons. For obvious reasons, this last form of assessment is linked with what is known as **mastery learning**.

In one sense, the standard or level of competence is really a pass-fail line, but the emphasis is different. There is no set number who can "pass". **Every** learner who shows the ability to perform at the required level is credited with the unit. Note again that quite commonly this means the person being assessed has to perform well in a range of things, and not just a single skill.

Two issues are worth considering here.

**How do I know when to use competency-based assessment?**

The answer is not as easy as we might want. In general, we choose a competency-based approach when it is important for the learner to demonstrate that they can competently do/know something in which a specific standard is required, rather than one in which a measure of a range of achievements is appropriate.

Driving a car is once again an obvious example. We simply cannot risk having learners driving alone if they can competently use the accelerator and the gears, but cannot also indicate, or brake at the right time.

Similarly, if the aim of the interpreting course discussed earlier is to provide interpreters to accompany business people trading in Arabic-speaking countries, then a high-level competency-based measure would be chosen.

If, however, an introductory interpreting course was provided for general and social use, then using a competency-based assessment programme would probably not be desirable, and certainly not be necessary. A sensible curriculum planner would realise that for assessment in such a course to be a rewarding part of the programme, learners should receive information about their level of performance, rather than be told that they have not reached the level of competence.

**How and where should the competency standard (or criterion) be set?**

Standards are sometimes unwittingly set at very high or low levels, often because the course goals are themselves unclear or unrealistic. The ability to be clear about where and why standards are to be set, is of critical importance in developing good competency-based assessments. This is not easy, and those who develop units and assessment programmes need to take this difficulty seriously.

If, for example, the competency level for getting credit in an elementary unit on carving is set so high that only a master carver could pass, then that would be counter-productive. At the same time, the use of a merit level can encourage learners to strive for excellence.

## ACHIEVEMENT-BASED ASSESSMENT

Despite its name, this is **not** the only form of assessment to focus on "achievement". All standards-based methods have individual learner achievement as one of their main concerns. In the Qualifications Framework, this particular term refers to the following:

> **Assessment in which a number of progressively more demanding standards are used; and in which all learner achievement is reported, usually in the form of a number or letter grade.**

Thus, in an assessment of weaving skills, a learner may perform in a very unskilled way. Nevertheless, what **was** achieved would be recognised, assessed to see which standard or level of achievement had been reached, and reported.

Achievement-based assessment will be used in the Qualifications Framework for most sixth and seventh form school courses and for programmes of general education. This is probably the type of **standards-based** assessment that teachers in secondary schools know best. The moderation trials and subsequent training programmes for Sixth Form Certificate used achievement-based assessment.

These sixth form trials used **grade-related criteria** as a way of arriving at an achievement-based assessment. In other words, learning outcomes were assessed by means of descriptions of what the learner had to do/know. These descriptions, called **criteria**, specified five or six levels of achievement for each skill or knowledge area. These levels of achievement were in turn normally linked with the grades 1 (low) to 5 (high). Then a formula was used to calculate what overall grade the learner was to be given for the year's work.

In French, for example, learners were assessed on their listening, speaking, reading and writing skills. For each of these skills, a set of criteria were established at five different achievement levels. The learner received the (numerical) grade for the level of skill they displayed in tests during the year. Then the scores for the four skills were used as a basis for determining a Sixth Form Certificate grade.

Thus a learner who could only extract fragments of information (the description or criterion statement), in a listening test would be given an achievement level of 1. A learner who could extract most information **and** cope with the unexpected gained achievement level 4 (cf. the extended example in Section Six).

Note that, just as students had to attain 50% to pass an examination, they will have to meet specified conditions to gain credit for a unit. The Qualification Authority's decision is that a learner must gain a grade 3 or better in at least half the skill or knowledge areas, and at least a grade 2 in the remainder.

There are several very positive aspects to an achievement-based approach.

- A **profile** of skills is often possible. The teacher can report back what the learner can and cannot do in all the skill/knowledge areas for which there are (assessed) learning outcomes.
- The various achievement criteria can be given to the learners, both to inform them of what is to be assessed, and as a means of helping them to set realistic goals.
- The criteria or descriptors attached to the various levels provide useful information to the learner and to others about what the unit is about, and what successful learners will be able to do.

At the same time, it should be recognised that there are some problems in achievement-based assessments.

- It is not easy to write clear and unambiguous descriptions of each grade level. In theory, each grade level should be more advanced and/or more demanding than the previous one. They should also be clearly distinguishable so that assessors can decide reliably which grade to award.

  In some cases, it may be difficult to avoid using words like "some", "most", and "all". This leaves to someone else the hard work of deciding what these terms actually mean in practice. Where there is a clear progression of difficulty, this vagueness might be possible to avoid, but writing criteria is something unit writers and teachers will need to work on very carefully.

- A different problem is deciding whether the grade achieved in a test given early in a course should count equally (or at all?), to similar tests given at the end of the unit. Teachers tend to have sharply differing views on this. Some feel that achievement during a course should be recognised, others believe that the final or "outcome" achievement is the only one that should count.

  In practice, it should depend on the nature of the unit content. If the learning outcomes describe several identifiable skills or material to be learned, then assessment of each of the parts seems a sensible idea, although the problem of how these assessments are combined remains. If the learning is progressive, as in the listening skills example noted above, then it seems more defensible to say that only an end-of- unit assessment should count.

  Where assessments do have to be combined, there is a further issue. How do we decide whether some outcomes are more important than others and, if so, should some grades or assessments be weighted to reflect this?

- For many school subjects, the syllabus consists of a variety of themes, blocks of content and/or skills which are either quite independent, or for which there are no clearly agreed levels of difficulty. This creates a further problem if the overall assessment is in terms of credit (i.e. pass/fail) for the unit, and the assessments have to be combined. Note that this is much less of a problem if **reporting is by a profile of results covering all parts of the unit.**

One way around the problem of a single grade is to break the subject up into units which **are** more coherent, and to assess each unit separately. This is one approach being advocated in the Qualifications Framework. Where teachers do not believe this is possible or desirable, the way in which an overall assessment is calculated does have to be seriously addressed.

## CONCLUDING REMARKS

Different forms of assessment have different goals. For good assessment programmes, **all** assessments require teachers to be aware of the purpose of the assessment and the strengths and weaknesses of the form of assessment which is chosen.

# 5 SOME CRITICAL ISSUES

This section introduces several further issues in standards-based assessment. In some of these there is no agreed solution or "right" method of handling the issue. Yet in every case some sort of decision has to be taken by an assessor. Other issues might have been included. The current selection was made in terms of what teachers are most likely to face in developing assessment programmes.

## THEORY AND PRACTICE

Before looking at other issues, the whole question of theory versus practice needs to be addressed. This booklet would not be particularly useful if readers decide that the ideas are all very fine in theory, but in "the real world" **none** of the ideas can be put into practice.

Some readers may not like this being said, but an avoidance of the hard issues is why some of our assessments are so bad, from national examinations down to classroom and workshop tests. If assessment programmes are to improve, we must find the best balance possible between what **should happen ideally** and what **can be done in practice**.

**The starting point should always be the ideal.** The teacher should consider first what would be the most valid assessment programme for the purposes of the unit and its individual learning outcomes. Then issues of reliability, bias (see below), and other points should be considered. This is often a time-consuming exercise. It also means that teachers have to be reasonably knowledgeable and confident about assessment principles.

**After that, practical issues must clearly be considered.** A starting checklist could include the following:

- **How much time would the ideal test/assessment take**

  a to prepare;

  b to administer;

  c to score/assess;

  and how much time can I make available?

- **How much assessing is involved in the ideal?** Is this desirable, given the length and importance of this particular unit? What can be modified if there is too much testing?

- **What finance, equipment, room or other facilities do I need** for the ideal, and what is available? How close can I get, with a minimum of rearranging?

- **Do I personally have the skills to administer and assess in the ideal way?** If not, can I get someone else to work with me or even do the assessing? Otherwise, how can I modify the ideal, but keep as high a level of validity as possible?
- **How much is lost in moving from the ideal?** This will be a judgment on the part of the assessor, but a judgment that should be informed by knowledge about assessment.

## HOW MANY (AND WHAT TYPE) OF ASSESSMENTS DO I NEED?

This point has already been referred to earlier. It involves issues of principle as well as practice. It is very easy, however, for one answer to this question - as often as possible - to result in a programme where assessment is continually on the minds of learners and teachers. Yet this **is** sometimes the best answer that can be given. The key issue is to decide what is most valid - and possible - for a particular unit or programme.

The first point here is that a single assessment of anything is open to the charge that a result was due to chance factors operating on that day. We discussed this point in part under Reliability in Section Three. As with other issues discussed in the present section, there are different considerations to be looked at for competency-based and standards-based assessment.

**In competency-based assessment**, there are several things to think about.

- **Do the learning outcomes really require a series of things to be tested separately, or can they be tested by one "general" performance?**

Often, the answer will be the latter, provided this shows that the learner has achieved competence in the total task. It is not much use demonstrating, on separate occasions, that you know where to stand, how to throw a tennis ball in the air, how to swing the racquet, and where to hit the ball, unless you can put this together and serve correctly in a game.

- **How many times (and when) can you allow a learner to re-sit a test of competence?**

Remember that the Qualifications Authority see reassessment as important, and that assessment for success is always a principle worth aiming for. You will need to have some plan for reassessment, **and** a rationale which explains your plan to learners.

In some cases, retesting will be a practical issue, depending on staffing, laboratory or equipment restrictions. This may mean, say, that only two attempts are possible in practice. You do need to consider the practical effects

of having learners request and successfully complete an assessment part-way through a course. What will they do for the rest of the time? What effect will their success have on other learners? What are the alternatives, if two or three learners show clearly that they **could** complete early?

- **Is one successful performance sufficient to demonstrate competence?**

We currently answer Yes in the case of a School Certificate examination, even if the learner has failed previously. Here, we should look carefully at the goals of the unit, and its learning outcomes, before making a decision. It may be that for important or dangerous tasks you should think about assessing on more than one occasion before accrediting a learner as "competent".

- **Are the measures themselves subjective, to the point where it is desirable to have another person assess the learner on a separate occasion before competency is recognised?**

In theory, this should not happen often, as good competency-based tests of clearly defined learning outcomes should be lead to reasonably reliable assessments. Unfortunately, the world does not always work like that.

On this same point, note that even when two people assess, it is far from certain that they will come up with the same answer. A set of reliable and tried criteria will again be critical here. A practice session for the assessors is also a very good idea whenever this is possible.

**In achievement-based assessment**, the issues are slightly different.

- **Is one final (summative) test the best way of assessing, or should all assessments made in a course be added together to give an overall grade?**

This latter approach is the traditional way in which many teachers have operated over the years. Depending on the aims of the unit, it may still be the best way.

There are, however, some serious issues involved when marks or other scores are added together. This is assuming the result is then to be translated into a grade level of some kind which, in turn, will be used to report on the learner's overall achievement. Teachers using such an approach should read more extensively on this, as the details are beyond the scope of this introductory booklet.

Note again that, for general education units (the kind schools mainly teach), the Qualifications Authority has already decided on a formula for combining achievement levels for overall credit in a unit (cf. Section Four).

- **Where multiple assessments are used, should there be equivalent tests on the *whole* learning outcome, or several tests of each *part* of that learning outcome?**

If several equivalent tests are given, is this a valid way of assessing performance? Does the learning actually build up in some way, so that the last score(s) should be used to determine the achievement level, but the earlier ones should not?

If on the other hand on-course tests are used, and they each test sub-skills, can using these scores be justified if the learning outcome is a single complex performance? And if the grade-related criteria for one element of a school course contain several assessable "bits", can we assess these separately and add the results together?

Remember here that **a profile of achievement can and should be given to the learner, even if the unit or qualification requirement is a single grade.**

## An Example

Both of the above issues can be illustrated by the following example. A teacher presents a six-week course on computing, in which the major learning outcome is the ability to work with a specified word-processing package. Suppose on the one hand that the teacher gives a global test at two-week intervals, and then averages the results. This automatically penalises learners who may have taken longer to come to terms with the programme, but who can work competently with the package by the end of the unit.

Suppose, on the other hand, that the teacher tests basic skills like paragraphing at the end of week one, shifting blocks of text after week two, and so on, then adds the scores together to give the final result. This may completely overlook the fact that a learner can do each bit of the unit, but is quite lost when set a global test of competence. The final result would also not be a valid measure of the major learning outcome noted above.

A third possibility would be for the teacher to give the same series of sub-tests, and a single global test at the end (does that ring bells?). What justification would/could there be for combining these scores? What would the teacher do with a learner who failed all sub-tests and showed competence in the final, global test? One answer would be to award the grade achieved in the final test and discount the rest. Another possibility would be to retest, just to make sure the single global result was not due to chance.

Sometimes these issues can be resolved by a more careful analysis of the learning outcome and the goals of the unit. It would not be valid for sub-tests to be used when a complex skill or performance is the goal; it may well be when the learning outcome can fairly be taken to mean competence in a series of sub-skills.

It is accepted that this particular discussion has raised a number of questions without providing many answers, but awareness of the possibilities may allow teachers to come to the "best" answer for a particular learning situation.

## MERIT PERFORMANCES

The awarding of merit is an important issue in standards-based assessment. It is easy to think of "credit" as equivalent to "average" - even though this is not how it is conceived - and merit awards do signal that excellence is still valued. Merit has been widely discussed in standards-based assessment overseas. Such awards look to be easier in **achievement-based assessment**. In one sense, this is true. We can simply say that people who achieve above a particular grade level, or who score a particular combination of high grades, are awarded a merit pass in the unit.

The Qualifications Authority has decided that **merit in achievement-based units for general education** should be awarded to learners who gain at least a 4 (of a possible 5) in all elements (aspects) of a unit.

Nevertheless, there are several unresolved issues in the award of merit in **competency-based programmes**, where no overall decision has been thought desirable. For this reason, a further booklet deals with merit for competency-based programmes.

It may be interesting simply to note that merit performances in achievement-based assessment programmes might be awarded on a similar basis to that in competency-based measures, rather than by a fixed formula. In developing achievement-based programmes for units where no required formula exists, teachers might find some useful ideas in the booklet referred to above.

## TEST DIFFICULTY

Test and measurement experts have been talking about the effects of test difficulty for years. The points to think about here are quite simple; what to do about them is not so easy.

If a decision is made in **competency-based assessment** about the level of skill or knowledge that is needed, then it would seem as if the "difficulty" of a test of that competency is irrelevant. Either the learners can demonstrate competency, or they cannot.

If, however, a physical education teacher is testing whether a learner can high jump 1.83 metres to qualify for a competition, it **seems** as though all that is needed is for the learner to jump that height, regardless of the "test" or the "test conditions". Yet it is clear that requiring **competition conditions** would provide a more difficult test than simply asking the learner to demonstrate in a practice session. To avoid this sort of problem, **the test conditions for competency have to be specified clearly.**

The situation is more complex in **achievement-based assessment**. The standard you set for each grade level may be somewhat arbitrary. What is also true is that, unless very carefully monitored, many of the tests of those levels may be even more arbitrary in their level of difficulty.

Suppose a teacher decides on an assessment task to determine map-reading skills in geography, where the criterion for a high grade is the ability to negotiate an outdoor course with the aid of a map. Issues like the time allowed, the nature of the terrain, even the size, clarity and colours of the map will affect the difficulty level of the test, and the chances of learners achieving at a high level.

The point here is that changing a few minor conditions in the test may result in a significantly higher or lower number of learners achieving particular grades. Once again, the test conditions should be clearly specified.

The issue of difficulty also arises in tests with a number of items, where the difficulty of **every item** will affect the distribution of grades achieved. This can lead to serious problems in reliability when two supposedly equivalent versions of a test have items of different difficulty (cf. Section Three).

## BIAS

**Bias always exists in assessment**, even though, as professionals, we would like to think our assessments - whatever else they may be - are unbiased. The issue of bias was noted in Section Two. It is mentioned again here, even though space prevents full and satisfactory treatment. The effects of bias can be serious and for some groups potentially devastating.

An assessment always has a frame of reference. It comes from a culture which assesses in a particular way. This is not necessarily "bad" or "good", but it means that a form of bias is inevitable. The common practice of assessment by oral examinations in some countries is not right or wrong, it is just how it is done. Being aware that we sometimes do things simply because That's the way we do it round here, is a good start, but we should try to ensure that such practices do not disadvantage groups we want to treat fairly and equally.

Three common forms of bias to watch for are as ethnic, gender and personal bias. It should be stressed that only one or two very limited points are offered in what follows.

- **Ethnic bias**. If all the examples, items or whatever in a test use only pakeha names, pictures, etc. there will clearly be bias. A more difficult area is the challenge to assess in a group rather than by individual performance. The learning outcomes need to be examined clearly to see what is possible, and that is where change may have to be negotiated. A biased assessment may be quite **invalid** for some groups.

- **Gender bias.** In a way similar to ethnic bias, if only males are used as models, there will be bias. You should also avoid stereotyping males and females, and never use sexist language. For example, do not use man or he, when you mean all people. In an age where such things are well-known, it is surprising how often they are ignored in practice.
- **Personal bias.** This can clearly take two forms. The "halo effect" occurs when a teacher either likes a particular learner, and/or believes that they are particularly competent. In both cases, it is very easy to inflate the scores given, or to give such a learner the benefit of the doubt in marginal decisions. On the other hand, a disliked learner, or one who has been lazy, misbehaves or just **seems** incompetent, is often marked down. Be aware of your personal prejudices in assessing. Use a second assessor if you realise you have strong feelings about a learner.

In all three cases, remember that bias in assessment is only one part of the story. Teachers should be careful not to allow bias to creep into their selection of curriculum or content material, or into their teaching practices. In both cases, even an apparently "fair" assessment will simply reflect such bias.

There **are** ways of preventing some forms of bias in assessment, but a critical starting point is being aware that it exists.

# 6 ASSESSMENT EXAMPLES

This section offers two extended examples in the field of standards-based assessment. **These examples are provided for the basis of discussion, and as an aid to further understanding of the issues. They are not presented simply as appropriate models in some cookbook fashion.** The second example, however, does suggest a series of steps that might be taken to develop an assessment programme.

The first example is drawn from a general unit of learning, located in a senior secondary school programme. It concerns assessment in a foreign language unit. Discussion in this example centres on the skill of listening.

The second example is drawn from the field of retail sales of goods and services. It provides discussion based on the draft of a unit standard in which - to date - no decisions on assessment have been offered. Once again, there is some selection of content, with the focus being on the ability to demonstrate customer handling skills.

## EXAMPLE 1: ASSESSMENT OF LISTENING IN A FOREIGN LANGUAGE UNIT

This example is based on a **draft** unit standard for sixth form foreign language study. It is important, therefore, to stress that the approved unit may look quite different.

**What is the unit, and how does Listening fit in?**

The proposed unit is a Qualifications Level 2 unit, worth 6 credits. It is one of four such units which, together, would typically make up a year's work in a foreign language. The unit is provisionally entitled, *Taking a Break in the Target Language*, and would cover such topics as travel, holidays, sport, media, eating, entertainment, cultural life and music.

The **elements or learning outcomes** relate to culture, listening, speaking, reading and writing.

**What are the learning outcomes for listening?**

For the listening skill, the draft specifies the learning outcome as follows.

At the end of this unit the student will be able to:

- **Listening**: listen to and analyse information heard in the target language about recreational situations and respond appropriately.

**What additional material is offered/available to the teacher?**

Along with the specification of elements, the teacher would be expected to use all of the following material.

- The present Sixth Form Certificate language syllabus, which includes sections on assessment.
- A statement of Achievement Criteria as part of the unit standard (i.e. the formal description of the unit).
- Three pages of Notes on the achievement criteria, giving suggestions on types of assessment

**What guidance is offered on assessment procedures?**

The statement of Achievement Criteria contains a preamble with three points for the teacher to note.

- A student who does not meet the minimum requirements of the Achievement 1 criteria will be awarded "Achievement 0" for that assessment.
- Even if not stated explicitly, achievement of the previous achievement(s) is assumed in each descriptor.[5]
- For each assessment task teachers need to develop an assessment schedule which makes explicit the relative terms "some", "most", and "all".

For the Listening skill, there are five levels of Achievement Criteria. This is likely to be common in many school-based units. The criteria (descriptors) are as follows.

- Extracts fragments of information.
- Extracts some basic connected pieces of information.
- Extracts most information.
- Extracts most information and makes inferences.
- Extracts all information and makes inferences.

The notes on these criteria intend to make what is to be done reasonably explicit. Because any discussion on what the teacher eventually does depends on these notes, they are in turn given here in full.

**Listening**

Possible sound sources for listening passages for summative assessments include the teacher, native speakers, and tape recorded or video-taped material. The passages may be spoken by a single person, or involve a conversation between two or more people.

A range of material should be used. Care should be taken to ensure that the technical quality of recordings is adequate and that the speed of delivery is appropriate.

5 *A "descriptor" is the statement which describes the criterion/criteria which must be met for the award of that grade/achievement level.*

**Language levels and content will be appropriate to the topic or topics upon which the assessment is based.**

An assessment task will involve listening (a conversation, a narrative, an advertisement, some instructions etc.) and carrying out one or more of the following activities:

- writing answers to questions
- extracting key points
- selecting from a possible choice of answers
- following instructions.

Answering objective-type questions (e.g. multi-choice, filling in gaps) may also be used, but not as a total assessment task. If this is used, another activity should be included in the assessment task as well.

Responses may be in English or in the foreign language, but if in the foreign language care must be taken to ensure that it is the listening skill that is being assessed.

**What form of assessment should be used?**

The writers of the draft unit seem to assume that teachers will use achievement-based assessment techniques. This is confirmed by other documents, which list the combination of achievement levels necessary for the whole unit to be credited to the learner.

Is this the best form of assessment in this case? Perhaps, but teachers should consider why, when there are specified requirements from an "external" constraint, namely, the *award of the unit*.

It seems reasonable to argue that the learner must reach some sort of competency level for us to award this unit. The Qualifications Authority has decided that in units like this, the student must reach level 3 in half the learning outcomes (elements), and at least level 2 in the remainder. An alternative to this might be the use of competency-based assessment, and the specification by teacher experts of what a satisfactory competence level would be.

**Are the assessment suggestions valid?**

In terms of the ideal, listening activities in real life might be compared with what is offered here. The two lists might look something like this.

- Real life listening
- people talking with us
- people talking to us
- people talking around where we are

- Unit standards tasks
- listening to native speakers
- listening to the teacher
- listening to a conversation

- listening to radios, etc.
- listening to TV, radio
- listening to a video
- listening on the phone

The match does not seem too bad, although we may want to argue that most listening also involves speaking, and wonder why the skills are assessed separately. That is a point language teachers may wish to pursue separately.

What does matter here is the extent to which the teacher attempts to use more realistic situations, as opposed to those which are less frequent in real life. For example, the notes to the teacher about technical quality apparently ignore the necessity in real life to listen to conversations in an almost constant context of background noise and other talk.

A second issue where validity is important is the match between the achievement criteria and the learning elements. For listening, remember, the learner has to listen and analyse information heard in the target language about recreational situations and respond appropriately. Teachers, then, need to ensure that any assessment programme does measure in a way which will allow them to say whether the learner can or cannot do this.

Here, the practical problem of testing the listening skills of a class of learners clearly interacts with what may be more valid in an ideal situation. The suggested responses listed earlier (writing answers to questions, etc.), do not seem like a particularly good match with our real life listening situations, where our responses would tend to be one of the following.

- Nothing - we just hear what we hear, although we may store some information, ideas, a tune, for recall/use later.
- A conversational response, which may include a request for repetition or clarification, a conversation filler, or a response which is totally "appropriate" but in a different sense:

  Q. What are you doing this evening?

  A. Get lost!

- Going to do something, as the result of a spoken request.
- Writing something down, such as directions, a shopping list, notes in class, or a telephone number.

The teacher should consider this issue of validity very carefully when making up an assessment programme. Clearly there are activities which can be carried out in the school and which are more valid than others. For example, following instructions would seem to be more valid than selecting from a possible choice of answers, something we very rarely do as a result of listening in real life.

A final point here relates to the validity of the criteria themselves. If we are good listeners in our own language, do we really extract all information when we listen and make inferences? We probably **do** make inferences, but what sort of inferences do we make in normal life? Does this not depend on our knowledge of the speaker, the subject and the context? How can a teacher decide what inferences are appropriate for expert learners to make when they are listening to a tape of a native speaker talking about going to a concert?

**What issues relating to reliability need to be considered?**

Reliability can be looked at here in terms of the suggestions offered, and in terms of what the teacher could do. In terms of what is offered, the immediate concern is with the definitions of the relative terms (some, most, etc.), used in the achievement criteria.

The most reliable way to define these terms could be to determine standards based on numbers of errors, or some similarly quantitative measure. Note that the **length** of what they hear, however, would make a difference to the likely number of errors made!

If a quantitative measure is chosen, it is important that the learners get to know this in advance. On the other hand, the tighter the definitions, the more questions there would be about the validity of the measure. Why **should** a learner who makes one more error than another (and therefore receives a lower grade), be assessed as **qualitatively** different?

Even if we specify the criteria carefully, we know that a reliable assessment is one which gives consistent results. We can raise the reliability of a listening assessment by eliminating everyday background noise (which varies in volume), and raise it again by ensuring that all learners have been exposed to the same content material. The more we do this, the more we risk decreasing the validity of the assessment. Why? Because the act of listening becomes more and more artificial.

**After considering validity and reliability, how does one then set appropriate standards?**

Most teachers do know their classes reasonably well. They do know the standards which are possible or expected in their subject, and have some good ideas about how to test against these standards. An experienced teacher arguably holds and uses a covert standard which enables them to set assessments at an appropriate level. Nevertheless, teachers should always be cautious about the kinds of issues raised here, many of which pose threats to reliability and, especially, validity.

**Are there other issues to consider?**

One point here relates to both reliability and validity. There are clearly many instances in real life where the best a very good listener can expect to do is to

extract fragments of information. **Which** fragments could help to distinguish between a very good listener and a weaker one. A reliable and valid test would need to distinguish appropriately and consistently between such listeners.

Remember the point made above - that the performance could simply be due to one listener having a much better knowledge of the subject area. If you and I both listen to a conversation at a noisy party about good hotels in Barcelona, we may be equally good listeners, but if you were lucky enough to have been to the Olympics, you will almost certainly extract more useful information than I will.

There are a number of further issues about which space prevents a detailed discussion of, but which teachers might like to consider in the light of other sections in this booklet.

- How many assessments should the teacher plan to give for the listening skill?
- How many **different kinds** of listening activity should the teacher assess to be confident about stating one learner is a good listener, but another is not?
- Should tests held early in the unit count, or only the final assessment?
- In either case, how is the teacher to be sure that performance on a particular test represents a fair sample of the learner's behaviour?
- Given that many language students are girls, should we use a majority of female voices?
- What range of accents, speech styles etc. should be included to make the listening tests more valid? Should we include speech of children as well as teenagers and adults?
- Should we include some words or phrases the learner has probably never heard, and include in our assessment the way in which the learner responds to these additions?

## *Concluding Remarks*

What this first example aims to demonstrate is that there are a host of issues to be dealt with in setting up a valid and reliable assessment programme, **even when there is considerable guidance from the unit standards and from a published syllabus.** These issues are even more serious for teachers who are completely new to the system, or who know little about standards-based assessment. Clearly teacher education and on-going monitoring are **critical** in such circumstances.

## EXAMPLE 2: ASSESSMENT IN A UNIT ON SALES TECHNIQUES

The second example is taken from a **draft** unit in the area of sales. The unit is located in the field of Retail, Hospitality and Tourism, and is provisionally entitled *Sell Goods and Services - Advanced*. The unit standard has been approved in draft by a Standards Advisory Group (June, 1992). It has been placed at Qualifications Level 3, and is worth 3 credits. Learners in this unit would normally have completed a Qualifications Level 2 unit (Sell Goods and Services), prior to entry.

The unit contains three elements, each with several performance criteria. The three elements are as follows:

- Identify elements of a high performance sale
- Sell a high intensity product or service
- Demonstrate customer handling skills.

Discussion will centre on the third of these elements, the ability to demonstrate customer handling skills. It should be noted that, to date, the draft does not contain a guide to assessment or moderation procedures. The discussion which follows, therefore, is meant to be illustrative of the steps which teachers or tutors in a training establishment might go through to establish a worthwhile assessment programme for the unit. It presupposes nothing, and should be read with that understanding.

**What are the performance criteria for the third element,** demonstrate customer handling skills?

The performance criteria for this third element are listed in the draft as follows:

a. customers are given prompt attention and other customers are acknowledged appropriately when salesperson cannot give them immediate attention

b. customers with special needs are identified. Range: special needs includes physically, intellectually disabled, transcultural, elderly

c. strategies for anticipating and meeting the needs of the above are effectively demonstrated

d. angry, difficult, passive complainers, disruptive customers are identified and dealt with effectively using communication techniques for negotiation and conflict resolution.

**What is the best form of assessment?**

The first question to be asked in this case is whether credit for the unit ought to be assessed using achievement-based or competency-based assessment. The assumption should not be made that one or the other system is automatically "better" for industry or school-based courses. Although it is likely that a competency-based measure will be the choice of the National Standards Body, there would seem to be some good arguments in favour of an achievement-based measure.

This suggestion relates to the whole notion of "credit" for a unit, which is currently awarded on an all-or- nothing basis. In turn, this means that in both an achievement and a competency-based measure, there must be a level at which the learner is said to have succeeded in the unit ("passed"), and can thus be awarded the three credits available.

Yet it seems reasonable to suggest that learners could come out of a unit like this one with a good deal of understanding and ability in sales, even if they had not reached the required level for the awarding of credit. This could be reported quite easily using an achievement-based approach.

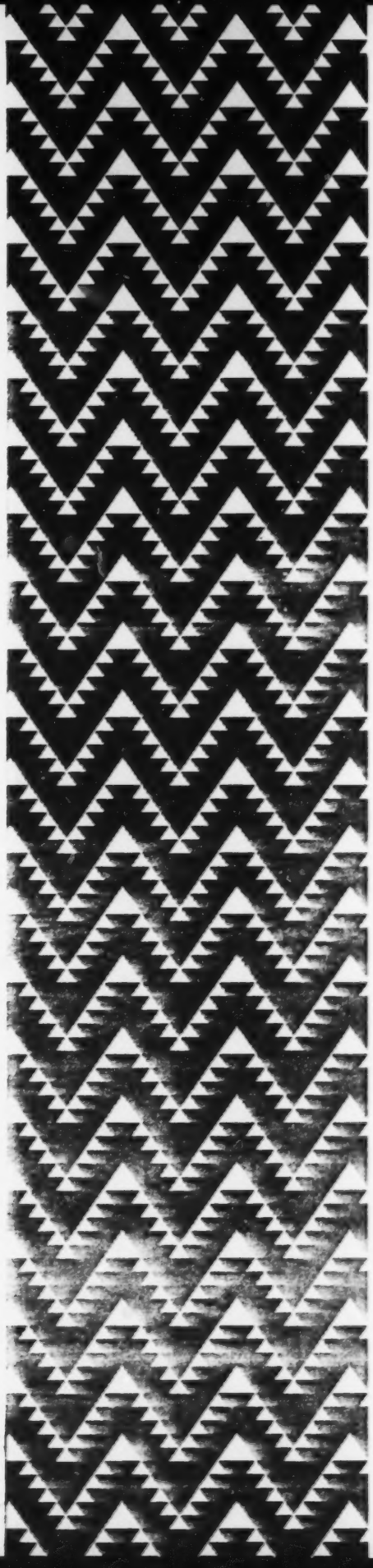
Assuming, however, that the assessment approach will be competency-based, there are still a number of questions to be considered. First, the purpose of the unit (specified in the unit standard), is that learners will be able to sell tangible goods and services of high intensity. High intensity is explained as goods and services in which the purchaser may have a low level of requirement and in which the purchaser invests a high level of interest. Purchases may involve hire purchase or finance agreements.

The intention, therefore, is quite specific. Learners who complete this unit satisfactorily will be able to **do** something, rather than just know it or know about it. The steps towards a worthwhile competency-based assessment programme for the customer-handling skills element of this unit might therefore proceed as follows.

**What is the ideal (most valid) assessment for this element?**

The answer is reasonably clear. The learner would need to demonstrate that in real-life sales situations they could satisfactorily meet all four of the performance criteria. A closer look suggests that this ideal would be extremely difficult to assess, if **all** parts of the performance criteria are to be observed in sales situations.

This is because various **combinations** of customer characteristics may occur, each requiring slightly different techniques on the part of the sales person. For example, presumably there would be a different technique employed in the following two cases.



### Case 1.

The sales person has to acknowledge appropriately [criterion (a)], a physically handicapped customer [criterion (b and c)], who is angry about an earlier deal and becomes disruptive .

### Case 2.

The sales person has to acknowledge appropriately a person from another culture who is clearly uncomfortable and upset, but quite passive.

The main difficulty, however, is that in a real-life sales situation the assessor might have to wait days or weeks for even the most common situations actually to occur.

**What valid assessments can be used which take practical concerns into account?**

This question is critical in the preparation of any assessment programme. Here, there might be a temptation to use role-playing situations, either in a real-life or in an appropriate mock-up of a sales area. This would certainly have higher validity than a written or verbal test in which the learner was asked what they **would** do. A written test, however, may well be **part** of an assessment programme, as it would allow for a much wider variety of situations to be covered than would be practically possible in most programmes.

To what extent would the role-playing situation be valid? This could depend on the way in which it was arranged. A role-play in a mock-up area and in which class members took the part of customers would be less valid than one in a real shop or office, and in which people unknown to the learner were used. A role-play in which several people of different types (age, gender, culture, other special needs), were involved, and where the sales trainee did not even know who were customers or who was with whom would add to the validity.

Ultimately, a standards body would need to decide on the extent to which a real-life situation was **necessary**, and the range or selection of customer situations that would have to be demonstrated in the various modes of assessing (written/verbal test, role-playing, real-life).

**Should the assessments be:**

- **specific, covering one part of one element;**
- **more general, covering all parts of an element;**
- **wide-ranging, covering all elements; or**
- **a combination of these?**

Note that this question is to do with the **assessment** and not the **teaching**. Again, issues relate to validity, but also to the reliability of the assessment.

An assessment programme in which tests are made of each (part of an) element clearly has the advantage of being more specific. This in turn means that observations of performance are likely to be more accurate, and therefore more **reliable**. At the same time, a salesperson in real life has to deal with **all** of the elements in some combination or other, so such a programme has lower **validity**. A Standards Body would have to decide what would be lost and gained here, and how important any losses might be.

**What opportunities should be given for repeated assessments, and what form should any such assessments take?**

This is an extremely tricky area. If, for example, on the first assessment a learner did not handle a difficult customer appropriately, what should a retest include? If the retest involves an almost identical situation, this could show whether the learner has in fact learned to deal with such a customer. On the other hand, any assessment of this kind is, remember, only a **selection from all possible situations**. To repeat **one** situation may be useful, but would it tell us how well the learner would deal with a different kind of difficult customer?

In this case, it might be more valid to retest by using a different scenario - assuming the scenario is under the control of the assessors. Clearly, a role-play situation would give this control, a real-life performance would not.

**How will competency be determined?**

This is arguably a matter for a Standards Board, composed of people from the sales industry, but informed by people who have some expertise in assessment. The **issues**, however, will be common to a wide range of units in courses aimed at industry.

Perhaps the main point to recognise is that, unlike areas where safety and/or health are involved, there is no clearly **objective** way of determining when a salesperson is "competent". Suppose, for example, a learner regularly and efficiently closes deals in real-life situations, but fails to meet many of the performance criteria for this unit. In this case, we would presumably say they did not meet the competency level. Yet is there something odd about that, as the purpose of the unit is to sell tangible goods and services of high intensity?

Suppose, on the other hand, that a learner demonstrated competence in all areas of the unit except one, but this one happened to be rather important all the steps of a sale are effectively completed. This is because the learner in question did everything "right", but somehow did not seem to effect sales. Would we be comfortable in saying they had achieved competency in this unit?

Undoubtedly a standards body would have to decide how many of the performance criteria would need to be demonstrated **and** under what conditions, but their selection would need to be based on what was considered reasonable for people working in this area. Awareness that the competency standard would be based in part on this experience, and **not** just on objectively defined sets of observable performance criteria, should help to determine the definition of a sensible and valid competency measure.

**Should there be an award of merit performances and how should these be assessed?**

As noted earlier, a further booklet is being prepared on merit aimed specifically at competency-based programmes. In brief, for a unit like this, the National Standards Body would first need to decide whether merit was appropriate for this unit, then on what basis it should be awarded.

The main difficulty would seem to be, once again, how the assessment could cover sufficient of the many possible sales scenarios to be sure that the learner was clearly superior. The issue here, then, is both one of **validity** (what is "meritorious"?), and **reliability** (how can we be sure that merit is deserved?).

**What special problems may occur in the assessment programme for this unit?**

Readers may like to refresh their memory on the problems of reliable observation at this point (cf. Section Three). This is one of the issues that would need to be addressed here.

A second issue is even more tricky. Sales often involve a "match-up" of the salesperson's technique and the characteristics of particular customers. This is, at least in part, an issue relating to human personality. A successful salesperson may be someone who has a particular personality, something not easily altered by the application of sets of sales "techniques". Yet different personality types may well be effective salespersons by applying techniques appropriate to their own personality.

Assessors in a unit like this would need to be very careful that they were actually **assessing the techniques specified in the performance criteria**, and not making judgments about the personality of the learner. Such judgments are in any case notoriously unreliable. In passing, it is worth noting that there is much the same difficulty in assessing teachers, as different personality types can lead to quite different ways of being "effective" in a classroom.

## A FINAL NOTE

What this discussion and the previous one illustrate is that assessment in **all** fields of education and training requires careful consideration of purpose, validity, reliability and practicality. It is true that in some units in industrial areas, factors of safety and/or health will dictate competency levels, but many of the other problems raised in these two examples will still operate.

Finally here, as valid assessment is closely linked to curriculum, both the Qualifications Authority and those teaching the units have a joint responsibility to ensure that assessment programmes are "fair, accurate and appropriate."

# 7 CONCLUDING REMARKS

Assessment is never easy, but it can and does affect people's lives. As teachers, we would do well to remember that. Standards-based assessment will be used in the Qualifications Framework. It is potentially a good way of assessing many areas of achievement; but teachers should be aware of the many thorny issues which can threaten the validity and reliability of any assessment programme. This awareness has been stressed throughout this booklet as a key factor to developing better assessments.

The booklet, however, must be seen only as a starting point, a brief introduction to issues in assessment. Even the most experienced of people working in assessment know that the **same** issues and the **same** problems come up each time a new assessment programme is developed. No assessment is going to be perfect, but we should strive to develop the best forms of assessment we can.

The starting point is to develop our assessment programme as an integrated part of our total curriculum plan. When our main purpose for assessing is clear, we should **choose an appropriate *form* of assessment, then select the most *valid* and *reliable* measures which are *possible and usable in practice*.**

Ma te wa ka kite
te pai o te hua.
Given time, you will see the quality of the fruit.
Kia ora koutou katoa.

# FURTHER READING


***Assessment For Better Learning: A Public Discussion Document.*** Wellington, NZ Department of Education, 1989.

Broadfood, P., Murphy, R & Torrance, H. (eds). ***Changing Educational International Perspectives and Trends.*** London, Routledge, 1990.

***Designing the Framework. A discussion document about restructuring national qualifications.*** Wellington, New Zealand Qualifications Authority, 1991.

Griffin, P. & Nix, P. ***Educational Assessment and Reporting: A new approach.*** Sydney, Harcourt Brace Jovanovich, 1991.

Hopkins, C.D. & Antes, R.L. ***Classroom Measurement and Evaluation.*** Illinois, Peacock, 1990.

***Learning and Achieving.*** Second Report of the Committee of Enquiry into curriculum, assessment and qualifications in Forms 5-7. Wellington, NZ Department of Education, 1986.

Gronlund, N.E. & Linn, R.L. ***Measurement and Evaluation in Teaching.*** (6th Ed.) New York, Macmillan, 1990.

Popham, W.J. ***Modern Educational Measurement. A Practitioner's Approach.*** New York, Prentice Hall, 1990.

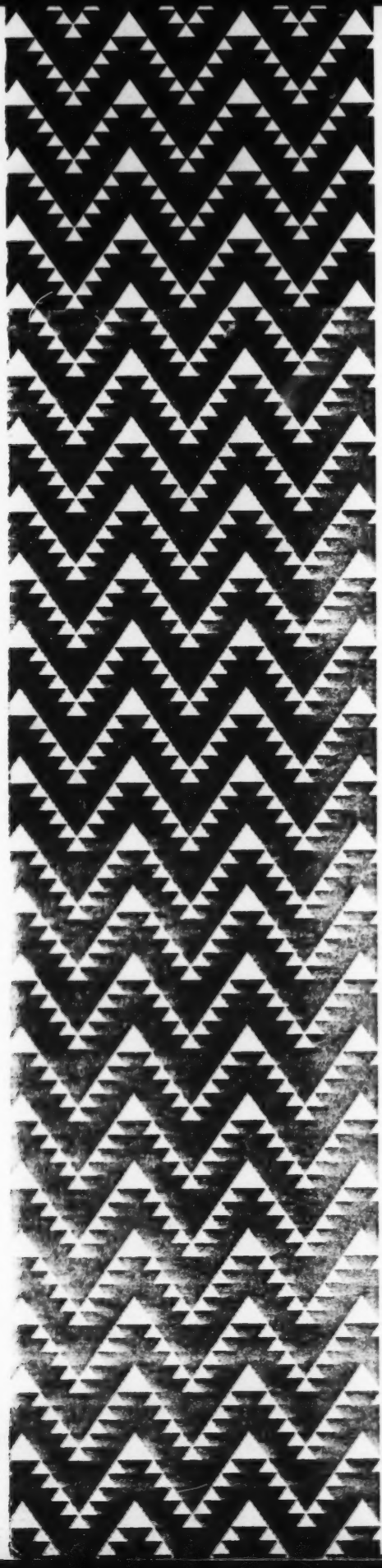
***Tomorrow's Standards.*** The Report of the Ministerial Working Party on Assessment For Better Learning. (The ABLE Committee Report). Wellington, NZ Ministry of Education, 1990.


**Notes**

The books by Gronlund and Linn, Hopkins and Antes, and Popham are standard texts on assessment.

Griffin and Nix is a recent (somewhat less detailed) book covering a range of assessment topics and issues.

Broadfoot et al. discuss trends in assessment, and not how to assess. The other works are recent New Zealand reports which discuss a wide range of issues relating to assessment.



NEW ZEALAND QUALIFICATIONS AUTHORITY
Mana Tohu Matauranga o Aotearoa

U-Bix Centre, 79 Taranaki Street, Box 160, Wellington, New Zealand
Phone: 04 385-0459 Fax: 04 385-4929

ISBN 0-908927-21-5